DIGITAL FUTURE
OR4 – Soluzioni Verticali

Analisi Servizi di Piattaforma per l'Healthcare

TR 4.10

AUTORE: EXPRIVIA S.P.A

Indice dei contenuti

INDICE DEI CONTENUTI **3**
INDICE DELLE FIGURE **4**
INDICE DELLE TABELLE **5**
1. INTRODUZIONE **6**
2. DESCRIZIONE DELLO SCENARIO **7**
3. SERVIZI DI PIATTAFORMA **10**
3.1. MACHINE LEARNING 10
3.2. ANALISI DEI DATI IN STREAMING DA DEVICE MEDICALI 11
3.3. CLUSTERING E ANALISI ESPLORATIVA 15
3.4. GESTIONE DEI DATI MULTIMEDIALI 16
3.5. DEEP LEARNING 17
4. CONCLUSIONE **44**
5. BIBLIOGRAFIA E SITOGRAFIA **46**

Indice delle figure

Figura 1 - Funzione di utilità per un paziente sepsi con tsepsis = 48 come esempio..... 9
Figura 2 - PC-300 Spot-Chek Monitor 12
Figura 3 - NuovoAir..... 13
Figura 4 - Xiaomi Mi Band 3..... 16
Figura 5 - Deep Learning per l'Healthcare 17

Indice delle tabelle

Non è stata trovata alcuna voce dell'indice delle figure.

1. Introduzione

Scopo del presente documento è descrivere uno scenario reale di utilizzo dei servizi esposti dalla piattaforma Digital Future in ambito Healthcare. Dalla descrizione dello scenario, si potrà evincere come, in più momenti diversi, la piattaforma Digital Future diventa uno strumento fondamentale per approcciare la grande quantità di dati che nel mondo Healthcare si raccolgono o vengono prodotti in real-time da dispositivi e attrezzature di ultima generazione.

Digital Future diventa, quindi, utile in più micro scenari, apparentemente distanti tra loro ma collegati da una analisi di fondo che a più riprese aiuta l'esperto del dominio in svariate attività: dalla raccolta dei dati, alla loro fruizione multimediale, all'analisi esplorativa, all'utilizzo di tecniche intelligenti per l'apprendimento da tali dati, all'analisi complessa di schemi "nascosti" tra i dati che vengono enfatizzati e portati alla luce.

Lo scenario che viene, di seguito, riportato enfatizza l'utilizzo pervasivo della piattaforma Digital Future utilizzando dati reali e ipotizzando come questi possano essere sfruttati per dar luce a differenti tipologie di analisi attraverso l'utilizzo dei servizi esposti dalla piattaforma.

2. Descrizione dello scenario

Un medico specialista in cardiologia, con una buona attitudine e capacità di analizzare i dati, dopo aver partecipato ad una conferenza sui rischi cardiologici legati alle sepsi, riceve un dataset con valori demografici e parametri vitali di pazienti, associati a valori di laboratorio relativi agli stessi su cui, per ogni record, viene associata una etichetta che indica valore 1 se in quel record era stata riscontrata la presenza di sepsi, 0 altrimenti.

Nella tabella seguente viene fornito l'elenco delle possibili informazioni (parametri vitali, valori di laboratorio e demografici*)* che sono presenti nel dataset ricevuto dal medico.

Dato	Descrizione
HR	Heart rate (beats per minute)
O2Sat	Pulse oximetry (%)
TEMP	Temperature (Deg C)
SBP	Systolic Blood Pressure (mm Hg)
MAP	Mean arterial pressure (mm Hg)
DBP	Diastolic Blood Pressure (mm Hg)
RESP	Respiration rate (breaths per minute)
EtCO2	End tidal carbon dioxide (mm Hg)
BaseExcess	Measure of excess bicarbonate (mmol/L)
HCO3	Bicarbonate (mmol/L)
FiO2	Fraction of inspired oxygen (%)
pH	N/A
PaCO2	Partial pressure of carbon dioxide from arterial blood (mm Hg)
SaO2	Oxygen saturation from arterial blood (%)

AST	Aspartate transaminase (IU/L)
BUN	Blood urea nitrogen (mg/dL)
Alkalinephos	Alkaline phosphatase (IU/L)
Calcium	(mg/dL)
Chloride	(mmol/L)
Creatinine	(mg/dL)
Bilirubin direct	Bilirubin direct (mg/dL)
Glucose	Serum glucose (mg/dL)
Lactate	Lactic acid (mg/dL)
Magnesium	(mmol/dL)
Phosphate	(mg/dL)
Potassium	(mmol/L)
Bilirubin total	Total bilirubin (mg/dL)
Troponinl	Troponin I (ng/mL)
Hct	Hematocrit (%)
Hgb	Hemoglobin (g/dL)
PTT	Partial Thromboplastin Time (seconds)
WBC	Leukocyte count (count*10^3/µL)
Fibrinogen	(mg/dL)
Platelets	(count*10^3/µL)
Age	Years (100 for patients 90 or above)
Gender	Female (0) or Male (1)
Unit1	Administrative identifier for ICU unit (MICU)
Unit2	Administrative identifier for ICU unit (SICU)
HospAdmTime	Hours between hospital admit and ICU admit

ICULOS	ICU length-of-stay (hours since ICU admit)

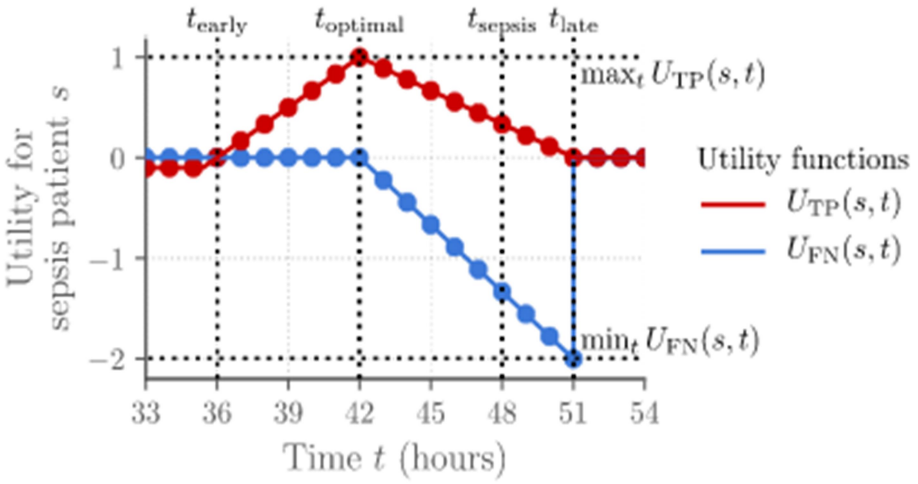


Figura 1 - Funzione di utilità per un paziente sepsi con tsepsis = 48 come esempio

3. Servizi di Piattaforma

3.1. Machine Learning

Il dataset ricevuto si configura come ideale per avviare una indagine di **Machine Learning**. Il medico specialista, che ha anche capacità di data scientist, procede a:

1. collegarsi e autenticarsi alla piattaforma **Digital Future;**
2. caricare il dataset ricevuto come file di tipo .csv nel file system HDFS gestito dalla piattaforma, utilizzando il servizio **Carica Files**;
3. analizzare i dati ricevuti tramite i servizi esposti dal componente **Data Governance** allo scopo di sviluppare algoritmi per l'apprendimento supervisionato.

La piattaforma, in maniera proattiva ed intelligente, supporta l'utente suggerendo una classifica ordinata di algoritmi di Machine Learning tramite il modulo **Algorithm Recommender** e grazie allo scouting tecnologico effettuabile tramite il componente **Mediator**. L'utente, quindi, riceve un feedback dalla piattaforma per approcciarsi al problema di apprendimento: gli algoritmi suggeriti sono valutati a seconda della loro efficacia nel contesto operativo di riferimento e della natura dei dati coinvolti. L'utente comprende che i primi algoritmi in classifica sono quelli con più probabilità di successo nel definire uno o più modelli da addestrare che siano adatti alla tipologia di dati in gioco. Per avviare l'esperimento di apprendimento con il dataset caricato, l'utente ha due possibilità:

1. Utilizzare il servizio **Apache Zeppelin** che consente di analizzare e mettere a punto un modello ML addestrato realizzando codice in Python. Il codice realizzato elaborerà i dati riversati in un **Dataset Offline** creato nel database Cassandra, a partire dal file precedentemente salvato nel file system HDFS attraverso una specifica funzionalità messa a disposizione dal portale.
2. Utilizzare il servizio **H2O** che consente, attraverso una procedura guidata, di analizzare e mettere a punto un modello addestrato a partire direttamente dal file caricato precedentemente nel file system HDFS della Piattaforma.

Una volta validato e testato, l'utente salva il modello realizzato nella piattaforma in modo che possa essere attivato durante l'analisi in real time dei dati in streaming dei pazienti.

3.2. Analisi dei dati in streaming da device medicali

In reparto arriva una nuova attrezzatura: il *PC-300 Spot-Chek Monitor* permette di trasferire digitalmente i parametri vitali raccolti da un paziente [1].

Lo *Spot-Check Monitor* è un dispositivo progettato per controllare i parametri fisiologici dell'utente, come la pressione sanguigna non invasiva (NIBP), la saturazione funzionale dell'ossigeno (SpO2), la frequenza cardiaca (HR) e la temperatura corporea (TEMP). Inoltre, il dispositivo è in grado di eseguire le misurazioni della Glicemia attraverso il collegamento della funzione Misuratore della Glicemia e i dati dell'ECG del Monitor Easy ECG (sia il Misuratore di Glicemia che il Monitor Easy ECG sono certificati separatamente). Questo dispositivo è applicabile per l'uso in istituti clinici e non ha condizioni o fattori di controindicazione.

Figura 2 - PC-300 Spot-Chek Monitor

- Le misurazioni accurate della pressione sanguigna possono essere attivate o annullate da un pulsante di scelta rapida.
- L'esclusiva tecnica di ossimetria assicura misurazioni rapide e accurate della SpO2 e della frequenza cardiaca tramite sensori intelligenti.
- La sonda di temperatura a infrarossi intelligente assicura misurazioni rapide e accurate della temperatura corporea.
- La pressione sanguigna, la saturazione di ossigeno, la frequenza cardiaca e la temperatura possono essere misurati simultaneamente.
- L'opzione Misuratore della Glicemia può essere collegata al dispositivo.
- È possibile contrassegnare fino a 100 ID utente.
- Memorizzazione dei dati con richiamo, fino a 999 gruppi di record possono essere memorizzati e riconosciuti tramite ID paziente.
- Gestione dell'alimentazione con modalità di risparmio energetico, spegnimento automatico e indicatore batteria scarica;
- Caricamento dati su PC tramite cavo USB e trasmissione dati in tempo reale su smartphone tramite connessioni wireless.

I dati della frequenza respiratoria vengono inviati invece da uno Spirometro intelligente, *Air Smart* o *NuovoAir* *[2]*.

Progettato per chi soffre di patologie respiratorie come l'asma, la BPCO e la fibrosi cistica, il dispositivo conduce, simultaneamente, il test della capacità vitale forzata e il test del volume espiratorio forzato. I risultati possono, da questi dispositivi, essere sincronizzati su un dispositivo intelligente o su un server cloud protetto, consentendo ai pazienti di condividere i dati con i loro team di assistenza.

Figura 3 - NuovoAir

Il medico o il suo supporto tecnico, decide di censire entrambi i device come dispositivi **plug-and-play** sulla *piattaforma Digital Future* e mandare in streaming i dati dei parametri vitali per controllare se tali parametri possono ricondursi ad un rischio legato alla sepsi.

Di seguito i possibili parametri vitali inviati alla piattaforma:

HR Heart rate (beats per minute)

O2Sat Pulse oximetry (%)

Temp Temperature (Deg C)

SBP Systolic Blood Pressure (mm Hg)

MAP Mean arterial pressure (mm Hg)

DBP Diastolic Blood Pressure (mm Hg)

Resp Respiration rate (breaths per minute)

EtCO2 End tidal carbon dioxide (mm Hg)

Una prima fase esplorativa dei dati potrebbe evincersi da una correlazione tra battito cardiaco (HR) e temperatura (Temp). Il medico, quindi, registra questi dati che arrivano a finestre di streaming dal PC-300 e utilizza la componente del **Correlation Analyser** della piattaforma Digital Future. Gli indici di correlazione confermano, ad esempio, una alta correlazione tra questi due fattori.

I dati in **streaming**, insieme ai dati dell'analisi del sangue del paziente che vengono inseriti manualmente dal medico costituiscono una tupla che può essere testata sul modello predittivo salvato precedentemente dal medico.

Con l'arrivo in streaming di questi valori sulla piattaforma Digital Future è possibile effettuare ulteriori indagini, ad esempio:

1. si può configurare un pattern complesso da individuare tramite la messa a punto di algoritmi capaci di gestire **Pattern di Eventi Complessi (CEP)** incapsulati in job java eseguiti attraverso il servizio **Apache Flink**;
2. si può avviare una fase di apprendimento non supervisionato per raggruppare i valori raccolti (**clustering**) secondo un comportamento ben preciso;
3. si possono avviare ulteriori indagini di **correlazione statistica (e semantica)** con i valori demografici ed eventualmente correlare tali dati ad un altro dataset.

Di seguito i possibili valori demografici associati ad un record:

Age Years (100 for patients 90 or above)

Gender Female (0) or Male (1)

Unit1 Administrative identifier for ICU unit (MICU)

Unit2 Administrative identifier for ICU unit (SICU)

HospAdmTime	Hours between hospital admit and ICU admit
ICULOS	ICU length-of-stay (hours since ICU admit)

Il medico, attraverso il servizio ***Apache Flink*** della piattaforma, imposta una query CEP memorizzata per l'individuazione di situazioni particolari che potrebbero portare il paziente ad eseguire ulteriori analisi mediche. Per questo tipo di attività è stata prevista una ontologia, memorizzata nel **Semantic & Ontological Layer** della piattaforma Digital Future, che ne descrive il significato patologico nel momento in cui la particolare situazione individuata dalla query CEP decide di inviare un messaggio di allarme al medico. Grazie ad una **query in SPARQL** generata dal CEP, l'ontologia viene interrogata con un ragionatore semantico per inferire un individuo ontologico (una regola) che descriva il significato ad alto livello del pattern complesso identificato dal CEP, come ad esempio l'evidenza di un rischio clinico derivato dalla serie di fattori patologici a loro volta inferiti da un altro individuo ontologico di più basso livello che ne descrive lo stato di gravità dagli intervalli medici noti in letteratura sui valori di laboratorio analizzati.

3.3. Clustering e Analisi Esplorativa

Al contempo, una fase di apprendimento non supervisionato può dare evidenza del raggruppamento di valori di laboratorio per tipologia di salute di un paziente. Il medico può utilizzare la sezione del servizio ***H2O*** dedicata al **clustering** di dati ed estrapola dei gruppi significativi che potranno essere utili alla comprensione di fattori comuni per la determinazione di una sepsi.

Il medico pensa che tutta questa mole di dati può essere analizzata e correlata con dei parametri di stile di vita e salute. Potranno essere utili un ulteriore dataset di dati storici riguardanti gli stili di vita e, in fase di esecuzione di un modello appreso, anche l'utilizzo di un *Fit Bit* (un braccialetto indossato da un paziente) per monitorare le abitudini giornaliere di frequenza cardiaca, calorie spese e ore di sonno medie. Uno *Xiaomi Mi Band* [3] è risultato un dispositivo censibile dalla piattaforma che può essere utilizzato per questo scopo.

Figura 4 - Xiaomi Mi Band 3

Il **Correlation Analyser** della piattaforma Digital Future è in grado di correlare tutti questi dati appartenenti a domini apparentemente differenti e presentare un risultato di correlazione statistica, spiegato semanticamente tramite una rappresentazione semantica latente in uno spazio vettoriale di ordine superiore.

I dati multivariati verranno proiettati automaticamente dal **Correlation Analyser** in uno spazio vettoriale ridotto per consentire la ricerca di un indice di correlazione. Questo risultato potrebbe eventualmente far sollevare dal CEP un allarme, se il medico decide di scrivere opportunamente una query di pattern complesso come eventuale risultato atteso secondo le sue intuizioni.

3.4. Gestione dei dati multimediali

Il medico riceve l'incarico di trasferire una serie di lastre digitalizzate in immagini su un supporto cloud come attività di backup. La piattaforma Digital Future, con il servizio di **Multimedia Manager**, si riconferma nuovamente uno strumento ideale, garantendo la persistenza delle immagini e la loro fruibilità in ogni momento.

Inoltre, il servizio di tagging automatico delle immagini consente di ricercare una immagine particolare a seconda del suo contenuto, estrapolato semanticamente dal

Semantic & Ontological Layer, e messo a disposizione del medico per semplificare e velocizzare la fase di ricerca delle informazioni associate alle immagini.

3.5. Deep Learning

La collezione di immagini raccolte e salvate sulla piattaforma potrà successivamente essere utile per attivare un task di object detection, un problema tipico nell'ambito del **Deep Learning**. La libreria di ***TensorFlow*** installata nella piattaforma può facilitare il medico nella realizzazione di questo problema di apprendimento supervisionato con un dataset di immagini.

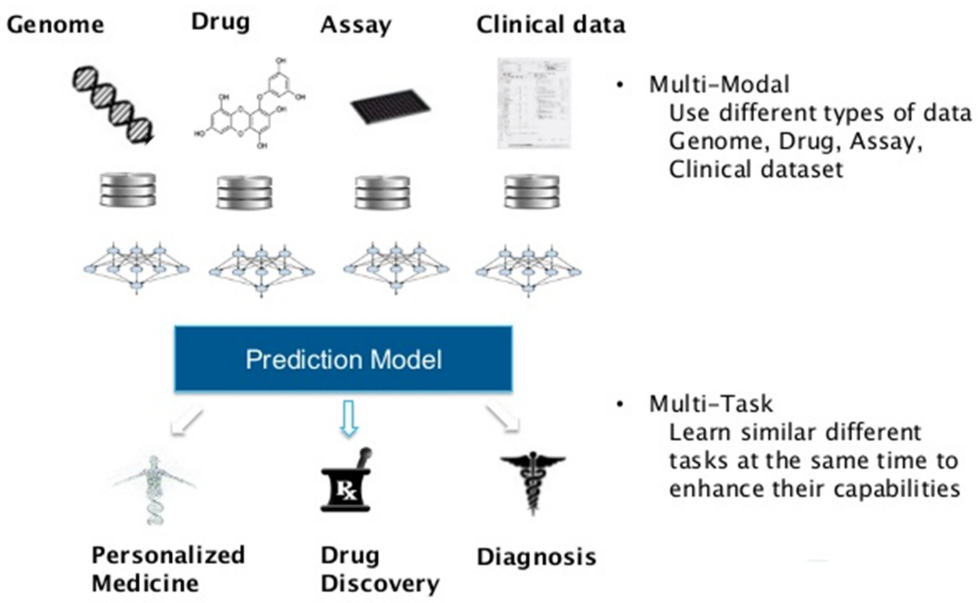


Figura 5 - Deep Learning per l'Healthcare

In generale, tutto il settore sanitario può ricevere grandi benefici dall'applicazione della data science e del deep learning nell'imaging medico. Le tecniche di imaging includono la risonanza magnetica (MRI), i raggi X, la tomografia digitale, la mammografia e così via. Vengono utilizzati numerosi metodi per affrontare la differenza di modalità, risoluzione e dimensione di queste immagini. Molti altri sono stati sviluppati per migliorare la qualità dell'immagine, estrarre i dati dalle immagini in modo più efficiente e fornire una interpretazione più accurata. Gli algoritmi basati sul deep learning che la piattaforma

Digital Future mette a disposizione aumentano l'accuratezza diagnostica apprendendo dagli esempi precedenti e quindi suggeriscono soluzioni di trattamento migliori.

Le tecniche di elaborazione delle immagini più diffuse si concentrano sull’enhancement (miglioramento), segmentazione e denoising che consentono un'analisi approfondita dell'anatomia degli organi e il rilevamento di diverse condizioni patologiche. Le applicazioni più promettenti mirano a rilevare tumori, stenosi delle arterie, delineazione di organi, ecc.

4. Progettazione di Scenari d'uso per l'utilizzo dei Servizi di Piattaforma Healthcare

4.1. Predizione del Rischio Clinico mediante Protocollo NEWS2

La possibilità di un monitoraggio continuo delle condizioni di salute rappresenta un aspetto cruciale per il miglioramento delle condizioni di vita, la prevenzione di potenziali patologie e la pronta risposta in situazioni critiche. In particolare, in terapia intensiva o situazioni di emergenza, la valutazione del grado di malattia di un livello di rischio clinico può essere considerata un compito predittivo in situazioni in cui flussi di dati di segni vitali vengono raccolti da dispositivi medici e sensori.

In questo caso d'uso, i sistemi di early warning che generano un punteggio aggregato basato sulla misurazione di un insieme di segni vitali, come il National Early Warning Score 2 (**NEWS2**), possono fornire una decisione utile supporto per la stima dello stato di salute e fattori scatenanti per interventi di terapia intensiva. In questo caso d'uso definiamo la tecnica di Machine Learning (ML) più adatta per la previsione delle classi di rischio cliniche di un paziente costantemente monitorato in una particolare condizione in cui è disponibile un numero limitato di parametri vitali. Questa analisi dovrebbe quindi essere preparatoria per l'obiettivo finale di progettare un dispositivo periferico collegato a uno o più dispositivi medici indossabili tramite IoMT, che sfrutta in modo adattivo il miglior modello ML per prevedere una classe di rischio NEWS2 affidabile.

Il protocollo NEWS2 si basa su un sistema di punteggio aggregato in cui un punteggio viene assegnato a misurazioni fisiologiche quando il paziente viene monitorato. Sei semplici parametri fisiologici formano la base del sistema di punteggio:

1. frequenza respiratoria (respiri al minuto);
2. saturazione di ossigeno (SpO2);
3. temperatura (°C);
4. pressione sistolica (mmHg);
5. frequenza cardiaca (battiti al minuto);

6. stato di coscienza, secondo lo schema AVPU. AVPU (Alert, response to Voice, response to Pain, Unresponsive) è un sistema attraverso il quale il personale clinico può misurare e registrare i livelli di coscienza del paziente.

Il protocollo NEWS2 è in realtà raccomandato dal National Health Service (NHS) del Regno Unito ed è particolarmente progettato per la determinazione di un grado di malattia di un paziente e pronto intervento di terapia intensiva nelle impostazioni:

- emergenza: per valutazione iniziale, monitoraggio e valutazione per il triage;
- monitoraggio: per la valutazione iniziale del ricovero e il monitoraggio continuo;
- pre-ospedaliero: per la comunicazione della gravità della malattia agli ospedali di accoglienza.

Nel protocollo NEWS2, scelto come protocollo target del caso di studio per testare il Modulo Analytics della Piattaforma di Telemedicina (cfr. Cap. 2 del PD13.3), un punteggio è associato a ciascun valore di parametro vitale rilevato, come mostrato nella Tabella 1. La somma di questi valori determina un punteggio finale aggregato, che viene utilizzato per delineare una classe di rischio clinico (basso-medio-alto) a soglie diverse. In particolare, solo per NEWS2, si dovrebbe aggiungere un punteggio di ponderazione di 2, inoltre, per ogni paziente che necessiti di ossigeno supplementare. Per NEWS2, la classe di rischio è:

- **basso**: il punteggio aggregato è 0–4;
- **medio**: se è 5–6;
- **alto**: se è ≥7 (fino a 20).

NEWS2 Vital Signs	3	2	1	0	1	2	3
Respiratory rate (bpm)	≤ 8		$9-11$	$12-20$		$21-24$	≥ 25
Oxygen saturation (SpO_2)	≤ 91	$92-93$	$94-95$	≥ 96			
Systolic blood pressure (mmHg)	≤ 90	$91-100$	$101-110$	$111-219$			≥ 220
Heart rate (bpm)	≤ 40		$41-50$	$51-90$	$91-110$	$111-130$	≥ 131
Temperature (°C)	≤ 35.0		$35.1-36.0$	$36.1-38.0$	$38.1-39.0$	≥ 39.1	
State of consciousness				A			VPU

Tabella 1 - Assegnazione punteggi per il calcolo della classe di rischio secondo l'algoritmo NEWS2

Qualora non fosse possibile rilevare tutti i parametri, si intende comunque procede con la fase di predizione del rischio clinico NEWS2 del paziente. In caso contrario, verrà eseguita la predizione sulla base dei parametri fisiologici registrati purché siano congrui

a quelli predisposti. Le combinazioni di misurazione previste per l'addestramento dei modelli di machine learning saranno le seguenti:

- 6 parametri vitali: frequenza respiratoria, saturazione di ossigeno, temperatura, pressione sistolica, frequenza cardiaca, stato di coscienza;
- 5 parametri vitali: frequenza respiratoria, saturazione di ossigeno, temperatura, pressione sistolica, frequenza cardiaca;
- 4 parametri vitali: frequenza respiratoria, saturazione di ossigeno, temperatura, frequenza cardiaca;
- 3 parametri vitali: frequenza respiratoria, temperatura, frequenza cardiaca.

La scelta di rimozione dei parametri vitali è motivata dalla necessità di poter dare comunque una predizione del rischio clinico nonostante l'assenza di alcuni parametri vitali che non sono stati rilevati o per mancanza degli opportuni dispositivi medicali di più difficile reperimento, come uno Sfigmomanometro, oppure per andare incontro ad esigenze temporali più ristrette di definizione di un early warning. Nel caso in cui non venga registrato il minor numero di misurazioni necessarie, il modello di machine learning non riuscirà a predirne il rischio clinico.

4.1.1. *Classificazione del Rischio Clinico mediante Apprendimento Automatico*

Predire il rischio clinico di un paziente, a fronte dei parametri vitali misurati e relativo calcolo del NEWS2 Score, è un problema di classificazione risolvibile con tecniche di Machine Learning di tipo Supervisionato in cui è possibile predire dati futuri, a partire dal Dataset iniziale aventi osservazioni ed output già noti, che supervisionano la bontà del modello.

Si dispone di un dataset, opportunamente anonimizzato, di pazienti sotto monitoraggio continui a cui sono stati rilevati i valori dei parametri vitali sopra citati.

Il dataset iniziale contiene, per diversi pazienti, le osservazioni dei sei parametri vitali misurati in una sessione temporale di circa sei ore. Lungo la sessione temporale, le osservazioni per ogni parametro vitale sono state misurate in maniera indipendente ed

identificate univocamente da un ID Sensore che verrà successivamente rimosso in quanto non contiene alcun contenuto informativo utile per la predizione dei dati.

id_assistito	id_sensore	cod_tipo_valore	valore	data
80	1	1	28	25/03/2019 08:30
80	4	2	10	25/03/2019 13:35
80	1	1	146	25/03/2019 14:10
80	8	4	38.08	25/03/2019 14:25
80	7	3	1	25/03/2019 08:35
80	1	0	113	25/03/2019 12:14
80	4	2	41	25/03/2019 10:13
80	8	4	38.49	25/03/2019 12:58
80	1	0	76	25/03/2019 13:26
80	1	1	68	25/03/2019 14:04
80	7	3	2	25/03/2019 10:30
80	8	4	34.95	25/03/2019 13:10
80	7	3	3	25/03/2019 13:41

Figura 6 – Stralcio del Dataset iniziale

Per ottenere dei modelli accurati, in grado di restituire una predizione affidabile, sarà adottato il tipico ciclo di vita di un'applicazione Data Science:

- Data Preparation,
- Data Exploration,
- Data Modeling.

Nella Data Preparation l'obiettivo principale riguarderà la ridefinizione del dataset principale in modo tale che per ciascun paziente venisse registrata la tupla contenente i sei parametri vitali, mediante l'utilizzo di tecniche di Time Series Analysis, e calcolare, una volta individuata la tupla completa, i valori NEWS2.

Una volta generato il Dataset X, conforme strutturalmente a quanto richiesto da un modello predittivo supervisionato, nella fase di Data Exploration saranno create copie del Dataset X in cui per ciascuno di essi, sono stati rimossi uno o più parametri vitali.

Nei Dataset così generati, saranno rimossi i seguenti set di parametri vitali:

- Dataset X_6 con 6 parametri vitali: presenti tutti i parametri vitali misurabili;
- Dataset X_5 con 5 parametri vitali: eliminato lo Stato di Coscienza;
- Dataset X_4 con 4 parametri vitali: eliminati lo Stato di Coscienza e la Saturazione;

- Dataset X_3 con 3 parametri vitali: eliminati Pressione Sistolica, Stato di Coscienza e Saturazione.

La fase di Data Modeling è un processo iterativo ed sarà applicato, senza alcuna distinzione, su tutti i Dataset X_n creati. La fase di Data Modeling è suddivisa in 3 step:

1. Campionamento del Training Dataset;
2. Hyper-parameters Tuning;
3. Comparazione dei Modelli

La fase di campionamento del Training Dataset ha il compito di estrarre dal dataset X un numero di sotto dataset e conformi matematicamente per l'addestramento del modello. Infatti, le osservazioni essendo temporalmente dipendenti, non garantiscono una predizione affidabile portando il modello in Overfitting. È necessario che i dati siano linearmente indipendenti. È stato definito un metodo che generi sotto dataset che rispettano i seguenti criteri:

1. Ogni dataset deve contenere m osservazioni indipendenti, associati ad m differenti pazienti. Ciò evita il problema di autocorrelazione, caratteristica tipica delle serie temporali;
2. Le proporzionalità delle classi NEWS2 devono rimanere preservate per ciascun dataset, in questo modo si prendono in considerazione le informazioni contenute nella popolazione dati di riferimento;
3. L'unione dei Dataset deve raggiungere almeno il 75% del Dataset X originale.

Il processo di campionamento genererà un numero imprecisato Dataset $D = \{D_1, D_2, \ldots, D_n\}$ e di Dataset $K = \{K_1, K_2, \ldots, K_m\}$ ottenuti dalla differenza tra il Dataset X e l'unione dei sotto dataset che compongono il Dataset D. I dataset D_n verranno utilizzati nella fase di Training del modello, mentre ogni dataset K è sottoposto ad Hyper-parameter Tuning. Nella fase di Hyper-parameter Tuning, si definiscono i migliori parametri per migliorare le performance del modello predittivo che, non sono ricavabili tramite dati e sono scelti dal Data Analyst. L'approccio utilizzato prevede la combinazione tra 5-Fold Cross Validation e Grid Search al fine di individuare i parametri che massimizzano la capacità predittiva. Per effetto del Grid Search, per ogni classificatore si individua un numero

finito di possibili valori che ciascun parametro potrebbe assumere e per ogni dataset appartenente a K poi si effettua il training dell'algoritmo utilizzando tutte le possibili combinazioni dei valori al fine di individuare quelli che massimizzano le metriche di valutazione. Nella 5-Fold Cross Validation, il Dataset viene suddiviso in k = 5 sotto Dataset dove i k-1 vengono utilizzati come Training del modello e il singolo dataset come Validation del modello. Il procedimento viene ripetuto per k volte utilizzando k combinazioni differenti.

I modelli di Machine Learning, che risolvono il problema della Classificazione, utilizzati e processati per Hyper-parameter Tuning sono:

1. Support Vector Machines (SVMs);
2. Random Forest (RF);
3. MultiLayer Perceptron (MLP).

I tre algoritmi sono i più diffusi e risolvono il problema della classificazione con tre approcci differenti. Nello specifico, SVMs riesce a risolvere problemi di classificazione non lineare proiettando dati non separabili linearmente in uno spazio a più dimensioni, trasformandoli i dati linearmente separabili tramite una Funzione Kernel quali Radial Basis Function (RBF) Kernel e Polynomial Kernel. RF rappresenta un ensemble di alberi decisionali, garantendo un modello più robusto i cui risultati non sono influenzati da normalizzazione, outlier e, di conseguenza, da overfitting. Infine, MLP è una tipologia particolare di Artificial Neural Network che utilizza l'algoritmo supervisionato di BackPropagation per l'addestramento e, tecniche di computazione non lineari per gestire features che non sono linearmente separabili.

I parametri utilizzati per ogni modello di Machine Learning sono:

1. SVMs: Radial Basis Function (RBF) Kernel e Polynomial Kernel;
2. RF: Numero tra 2 fino a 50 alberi decisionali;
3. MLP: Numero tra 1 fino a 100 Strati Nascosti e le funzioni di attivazione Logistica, Tangente, Rettificatore (ReLu).

Per ogni dataset e per ogni classificatore, sono stati scelti i migliori classificatori tra gli m dataset sottoposti nella fase di Hyper-parameter Tuning e denominati come SVM_{best}, RF_{best} and MLP_{best}.

Dai modelli SVM_{best}, RF_{best} and MLP_{best} verranno estratti i parametri migliori da utilizzare per ogni dataset D durante il Training dei modelli, classificando le classi di rischio NEWS2. Il Dataset D sarà quindi suddiviso come 80% Training e 20% Testing.

4.1.2. *Explanation dei Risultati di ML mediante Rappresentazione della Conoscenza e Ragionamento Automatico tramite Ontologie*

Un modello di ML supervisionato che risolve un task di classificazione sarà in grado di indicarci, a partire da un set di ipotesi, in quale classe di appartenenza vengono collocate le osservazioni del dataset. Tuttavia, la classe predetta non fornisce alcun ragionamento intrinseco e per questo è necessario che venga associata un'ontologia in grado di rappresentare la conoscenza e dedurre da essa, un ragionamento semantico. Un'ontologia può essere interpreta come una rete semantica in cui include l'insieme dei concetti (entità, attributi e processi), che ne definiscono il vocabolario di termini, gli assiomi e le relazioni tra concetti.

Si intende quindi costruire un'ontologia medicale di tipo Top-Down in cui sarà definito il livello Upper Domain Ontology, ovvero i concetti generali comuni per poi specificare i concetti che non possono essere decomposti nel livello Lower Domain Ontology. I sei parametri vitali rappresentano le entità di alto livello e per ciascuna di essi sono state definite più sotto entità, stabilendo una relazione di ereditarietà. La modellazione dell'ontologia medicale riprodurà fedelmente il sistema di punteggio assegnabile a ciascun valore di ogni parametro fisiologico, grazie all'utilizzo dell'Annotation Property "Valore Numerico" aventi le sottoproprietà "ValoreMin" e "ValoreMax" che hanno permesso la definizione delle sottoclassi per ciascuno dei parametri vitali. Per ogni Annotation Property è indicata una etichetta alle sottoclassi, affibbiate secondo il ragionamento derivante dal sistema di score del NEWS2:

- +3: "Estremamente Basso (Criticità)";

- +2: “Molto Basso”;
- +1: “Basso”;
- +0: “Normale”;
- +1: “Alto”;
- +2: “Molto Alto”;
- +3: “Estremamente Alto (Criticità)”.

Per lo stato di coscienza diviso è stato imposto la label “Vigile” con sottoproprietà 0, il quale indica la lettera ‘A’, e “Stato non Vigile” con sottoproprietà 3, il quale indica le lettere V (Reazione Vocale), P (Reazione al Dolore), U (Coma).

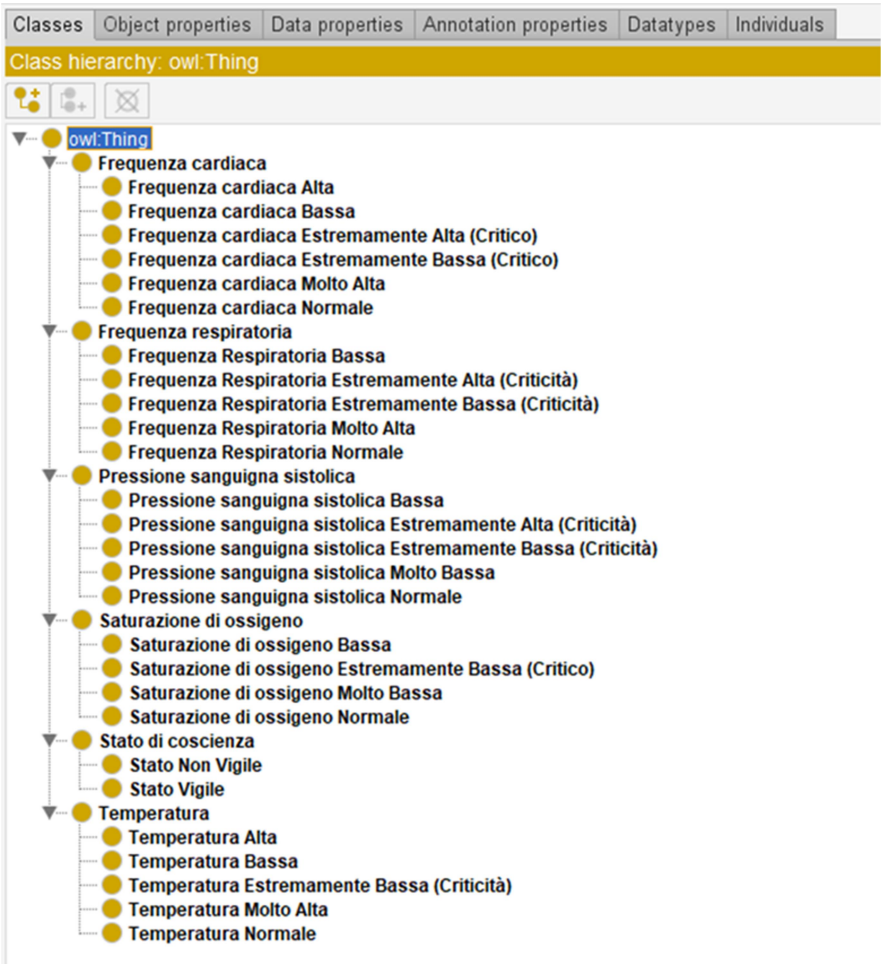


Figura 7 - Rappresentazione tassonomica dell'Ontologia di riferimento per lo Scenario d'Uso

4.2. Classificazione dei Referti Clinici

La crescita esponenziale di pubblicazioni scientifiche in ambito medico ha portato, in particolare nell'ultimo decennio, alla messa a punto di numerose tecniche di Information Retrieval (IR) finalizzate al reperimento di informazioni rilevanti e quanto più possibile attinenti a quanto ricercato.

Le tecniche di information retrieval basate sulla semplice estrazione terminologica contenuta nei testi, come ad esempio le tecniche basate esclusivamente sulla funzione di scoring TF-IDF (term frequency–inverse document frequency) utilizzata per misurare la rilevanza di un termine rispetto ad un documento o ad un insieme di documenti, si sono dimostrate estremamente limitate e non sufficienti al reperimento delle informazioni di interesse. Questo per lo più a causa dell'enorme complessità del linguaggio adottato in ambito medico, complessità dovuta non solo alla grossa quantità di specifici termini presenti nel vocabolario adottato in questo ambito ma anche a fattori quali l'eccessiva ambiguità con cui tali termini vengono utilizzati, e alla non "canonicità" delle complesse frasi adottate nelle pubblicazioni e nei referti medici.

I classici sistemi di information retrieval non sono infatti in grado di distinguere tra le strette correlazioni semantiche tra entità (ad esempio tra le differenti patologie cliniche) trattate in un testo e le relazioni tra entità determinate dalla semplice co-occorrenza dei termini di ricerca all'interno di uno stesso documento o paragrafo. Un medico potrebbe essere ad esempio interessato a capire se vi sono delle correlazioni tra le patologie al fine di comprendere il quadro clinico di un paziente o semplicemente reperire le informazioni circa i farmaci con cui un determinato disturbo clinico viene trattato e gli eventuali effetti collaterali. Un ricercatore in ambito biomedico potrebbe essere interessato a studiare le variazioni di un determinato gene, i percorsi che sono interessati da quel gene e le malattie che possono essere causate da quel determinato gene. L'utilizzo di motori di ricerca specifici quali PubMed[1] ad esempio non consentono

[1] https://www.ncbi.nlm.nih.gov/pubmed/

di reperire facilmente le informazioni ricercate: la conoscenza necessaria per studiare tali argomenti è presente ma il problema è come riuscire ad ottenere facilmente queste informazioni.

La presa di coscienza che la rappresentazione dei testi come semplici bag-of-words non consente di mettere in luce questo tipo di relazione tra termini, ha portato negli anni all'osservazione che un sistema di retrieval per poter essere efficace deve tener conto dello specifico dominio di conoscenza a cui afferiscono i documenti ricercati, essere in grado di collegare i documenti a tale dominio al fine di reperire le informazioni di interesse per l'utente con una priorità elevata e renderle infine facilmente fruibili all'utente.

4.2.1. *Tecniche per la classificazione dei documenti*

In particolare, nell'ultimo decennio, i ricercatori hanno cercato di migliorare la rappresentazione dei documenti attraverso una modellazione concettuale degli stessi. Varie tecniche di machine learning sono state adottate per mettere a punto tali soluzioni, una di queste si basa sull'utilizzo combinato di tecniche di **Information Extraction** (IE) e **Ontologie**.

4.2.1.1. *Le Ontologie*

L'ontologia, "scienza dell'essere"[2], tipicamente presenta differenti significati in base al contesto in cui si colloca. La parola ha origine nel campo della filosofia dove vari filosofi da Aristotele (4th secolo) a Leibniz (1646-1716) e più recentemente nel 19-esimo secolo studiosi di ontologie quali Bolzano, Brentano, Husserl e Frege hanno fornito criteri specifici per distinguere tra differenti tipologie di oggetti e relazioni tra essi.

[2] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," Sci. Am., vol. 284, no. 5, pp. 34–43, May 2001.

Nel 20-esimo secolo, l'intelligenza artificiale (AI) ha adottato questo termine inteso nello specifico come una "formale specificazione di concettualizzazione"[3] nel contesto della conoscenza e della condivisione dei dati.

L'ontologia può essere definita come un modello dichiarativo di un dominio che rappresenta i concetti esistenti in quel dominio, i loro attributi (o proprietà) e le relazioni tra esse[4]. Un aspetto essenziale delle ontologie è che devono essere formali e codificate in un linguaggio interpretabile da una macchina[5].

Le ontologie sono diventate comuni nel web, e variano da grandi tassonomie di dominio che caratterizzano siti web (quali Yahoo!), alla categorizzazione dei prodotti in vendita e delle loro peculiarità (ad esempio Amazon). Il Consorzio W3C ha codificato il Resource Description Frameword (RDF)[6], un linguaggio per codificare la conoscenza intrinseca nelle pagine web al fine di renderla comprensibile ai motori di ricerca. Molte discipline nel corso degli anni hanno sviluppato ontologie standardizzate che gli esperti di dominio possono utilizzare per condividere e annotare informazioni nei loro specifici campi.

---

[3] T. R. Gruber, "A translation approach to portable ontology specifications," Knowl. Acquis., vol. 5, no. 2, pp. 199–220, Jun. 1993.

[4] O. Corcho, M. Fernández-López, and A. Gómez-Pérez, "Ontological Engineering: Principles, Methods, Tools and Languages," in Ontologies for Software Engineering and Software Technology, Springer Berlin Heidelberg, 2006, pp. 1–48.

[5] F. Ruiz and J. R. Hilera, "Using Ontologies in Software Engineering and Technology," in Ontologies for Software Engineering and Software Technology, Springer Berlin Heidelberg, 2006, pp. 49–102

[6] Brickley, D. and Guha, R.V. (1999). Resource Description Framework (RDF) Schema Specification. Proposed Recommendation, World Wide Web Consortium: http://www.w3.org/TR/PR-rdf-schema.

In ambito medico, nello specifico, sono stati prodotti vocabolari di grandi dimensioni, standardizzati e strutturati come SNOMED[7], e la rete semantica dell'**Unified Medical Language System**[8].

Un'ontologia definisce un vocabolario comune per i ricercatori che devono condividere informazioni in un dominio e include definizioni interpretabili automaticamente di concetti di base nel dominio e relazioni tra loro. Ma qual è la reale utilità nell'utilizzo di una ontologia? Alcuni dei principali motivi sono i seguenti:

- Per condividere la comprensione comune della struttura dell'informazione tra persone e agenti software[9].
- Per consentire il riutilizzo della conoscenza del dominio.
- Per separare la conoscenza di dominio dalla conoscenza operativa.
- Per analizzare la conoscenza di dominio;

Le **ontologie** possono essere classificate in base al livello di conoscenza dettagliata che forniscono. Le ontologie upper-level forniscono una conoscenza molto generica con un basso livello di conoscenza specifica di dominio: rientrano ad esempio in questa categoria le ontologie delle patologie mediche (Disease Ontology) in cui sono definire tutte le patologie senza un particolare focus su uno specifico dominio.

Le ontologie generali rappresentano la conoscenza ad un livello intermedio di dettaglio, mentre le ontologie di dominio rappresentano la conoscenza relativa ad un particolare

---

[7] Price, C. and Spackman, K. (2000). SNOMED clinical terms. BJHC&IM-British Journal of Healthcare Computing & Information Management 17(3): 27-31

[8] Humphreys, B.L. and Lindberg, D.A.B. (1993). The UMLS project: making the conceptual connection between users and the information they need. Bulletin of the Medical Library Association 81(2): 170.

[9] Musen, M.A. (1992). Dimensions of knowledge sharing and reuse. Computers and Biomedical Research 25: 435-467

dominio e devono rappresentare tutti i concetti sottesi allo specifico dominio rappresentato: un esempio di ontologia di dominio è rappresentata ad esempio dall'ontologia relativa alle malattie genetiche. Infine, le ontologie designate a descrivere uno specifico task sono dette ontologie applicative.

Le figure che seguono forniscono 2 esempi di ontologie in ambito medico, una relativa alle relazioni che intercorrono tra patologie del sistema immunitario e l'altra che descrive invece un flusso operativo ospedaliero.

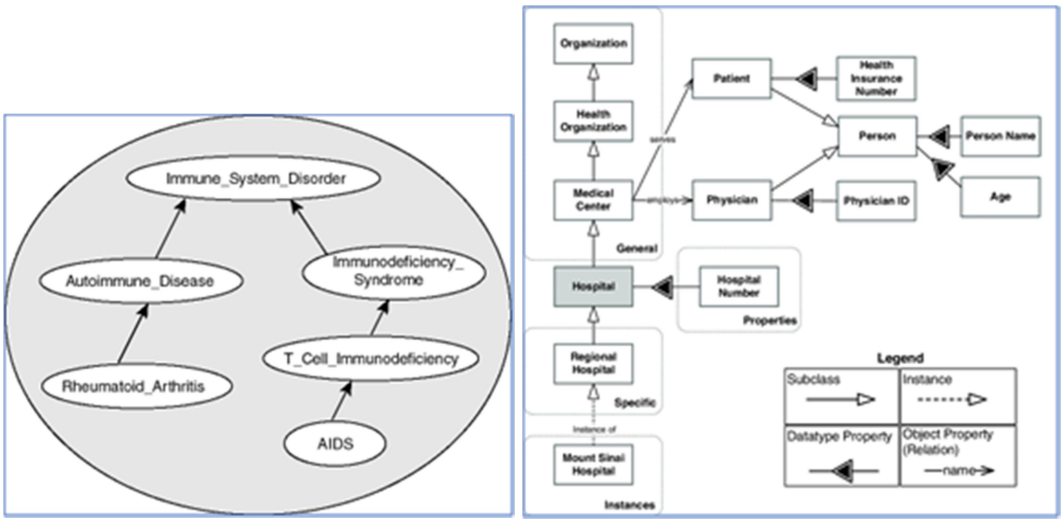


Figura 8 - (a) Ontologia relativa alle patologie del sistema immunitario; (b) Ontologia relativa ad un flusso ospedaliero

In particolare, osservando la prima figura, si evince che la diagnosi Immune_System_Disorder include due sottoclassi, Autoimmune_Disease e Immunodeficiency_Syndrome. La diagnosi Autoimmune_Disease comprende la sottoclasse Rheumatoid_Arthritis; e la diagnosi Immunodeficiency_Syndrome comprende la sottoclasse T_Cell_Immunodeficiency, che include la sottoclasse AIDS.

Nella sperimentazione descritta nel presente documento, l'obiettivo è stato quello di classificare referti medici in lingua italiana e in lingua inglese tramite delle ontologie di tipo upper-level al fine di facilitarne il reperimento attraverso un'applicazione web contenente apposite funzionalità di ricerca. La componente di classificazione dei referti è stata infatti integrata, come vedremo in dettaglio nel paragrafo "**Opensemanticsearch**", all'interno di un'applicazione open-source che è stata adattata

alle specifiche esigenze del progetto. Tale sperimentazione si colloca all'interno della piattaforma di telemedicina, sistema sviluppato da Exprivia, per il teleconsulto, il telereporting, la telepresenza e il telemonitoraggio per supportare tutti gli operatori coinvolti nella cura e nel monitoraggio del paziente a casa.

Per la sperimentazione descritta nel presente documento come vedremo più in dettaglio nei prossimi paragrafi sono state utilizzate 2 ontologie upper-level:

- l'ontologia **MeSH**[10] (**Medical Subject Headings Thesaurus**), un vocabolario controllato e organizzato gerarchicamente prodotto dalla National Library of Medicine, utilizzato per l'indicizzazione, la catalogazione e la ricerca di informazioni biomediche e relative alla salute. MeSH include i titoli dei soggetti che compaiono in MEDLINE / PubMed, nel catalogo NLM e in altri database NLM;
- la **Disease Ontology**[11] (**DO**), un'ontologia standardizzata per le malattie umane finalizzata a fornire alla comunità biomedica descrizioni consistenti, riutilizzabili e sostenibili relative alle malattie umane, caratteristiche del fenotipo e concetti relativi alle malattie inerenti il vocabolario medico. La realizzazione di questa ontologia è avvenuta grazie alla collaborazione dei ricercatori degli istituti di prestigio quali Northwestern University, Center for Genetic Medicine e University of Maryland School of Medicine, Institute for Genome Sciences. Disease Ontology integra semanticamente malattie e vocabolari medici attraverso un'estesa mappatura incrociata dei termini DO verso MeSH, ICD, thesaurus di NCI, SNOMED e OMIM. Nella presente sperimentazione, in particolare, è stata utilizzata la versione **DOID (Disease Ontology Identifiers)** dell'ontologia DO in cui ciascuna voce è identificata dal prefisso DOID seguita da

[10] https://www.nlm.nih.gov/mesh

[11] https://disease-ontology.org/

un numero: ad esempio l'Alzehimer Disease è rappresentata tramite un identificatore stabile DOID:10652.

Nello specifico l'ontologia MESH è stata utilizzata per la classificazione di re**ferti medici in lingua italiana** ripulite di informazioni sensibili quali anagrafiche e strutture di erogazione mentre l'ontologia DOID, è stata adottata per la classificazione dei **referti medici in lingua inglese** relativi al **COVID19** e reperiti dal sito della **SIRM** (Società italiana di Radiologia Medica e Interventistica)[12].

Entrambe le ontologie, rispettivamente MESH e DOID, sono espresse in formato **OWL** (W3C Web Ontology Language)[13], linguaggio per il semantic web[14] specificatamente progettato dal W3C per rappresentare ontologie, definendo esplicitamente significato e semantica di termini con vocabolari e relazioni tra gli stessi. Nello specifico i componenti principali di una ontologia OWL sono individui, proprietà, classi e valori dei dati. Gli individui rappresentano gli oggetti nel dominio di interesse, le proprietà sono relazioni binarie (ovvero che collegano due oggetti per volta) tra individui, le classi sono gruppi di individui.

Le figure che seguono mostrano una rappresentazione delle 2 ontologie utilizzate nella presente sperimentazione, ottenuta mediante l'utilizzo del tool Protege[15], un editor per ontologie open source specificatamente progettato per la costruzione di sistemi basati sulla conoscenza e ampiamente adottato in svariati campi dalla biomedicina all'e-commerce.

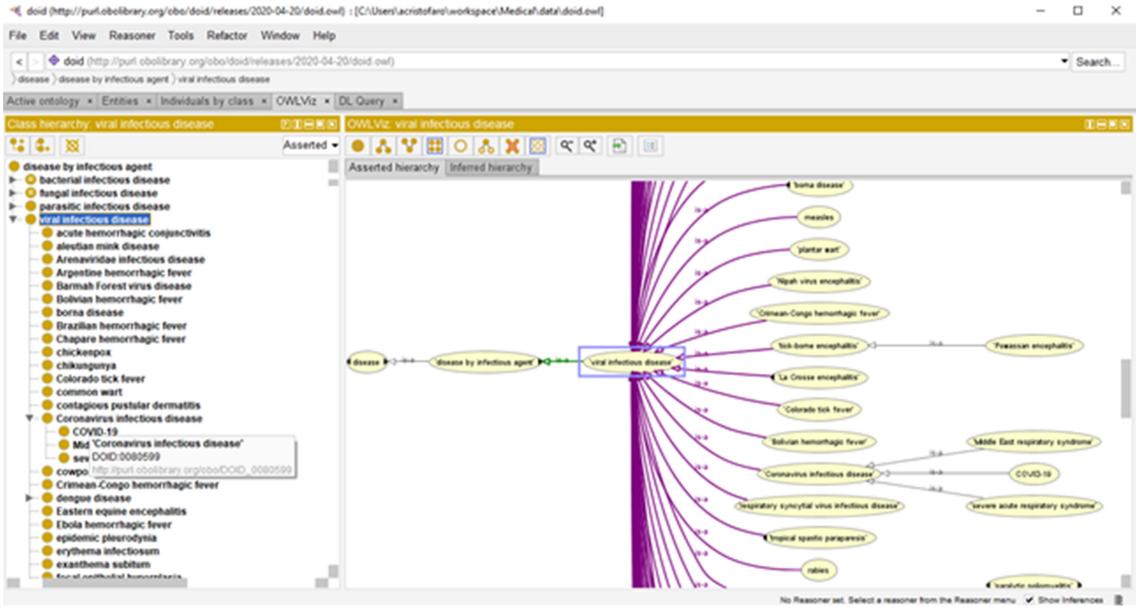
---

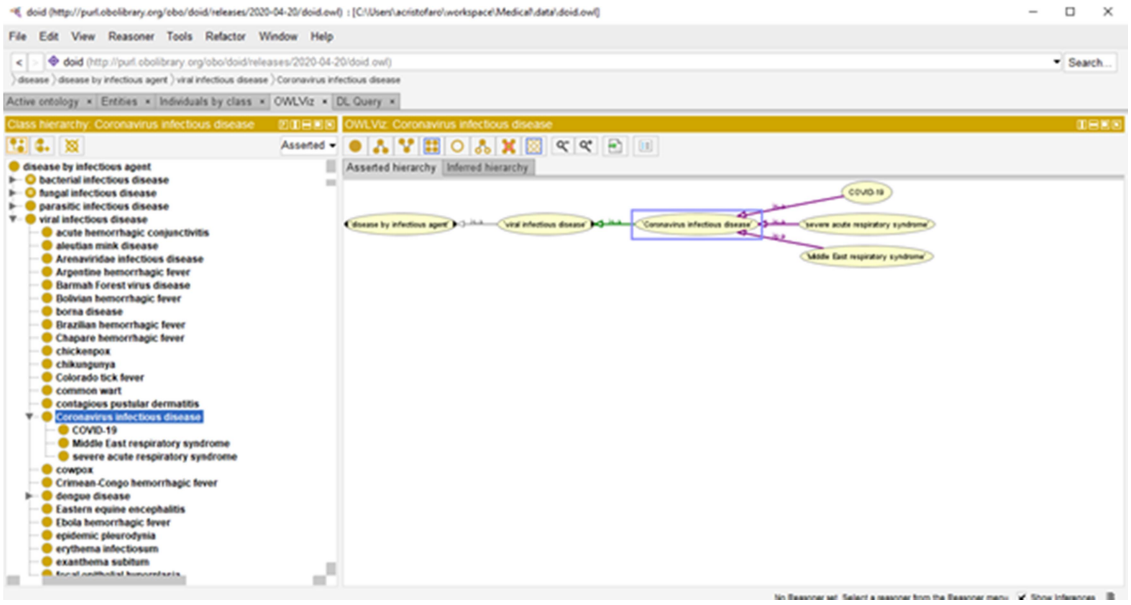
[12] https://www.sirm.org/

[13] https://www.w3.org/OWL/

[14] https://www.w3.org/standards/semanticweb/

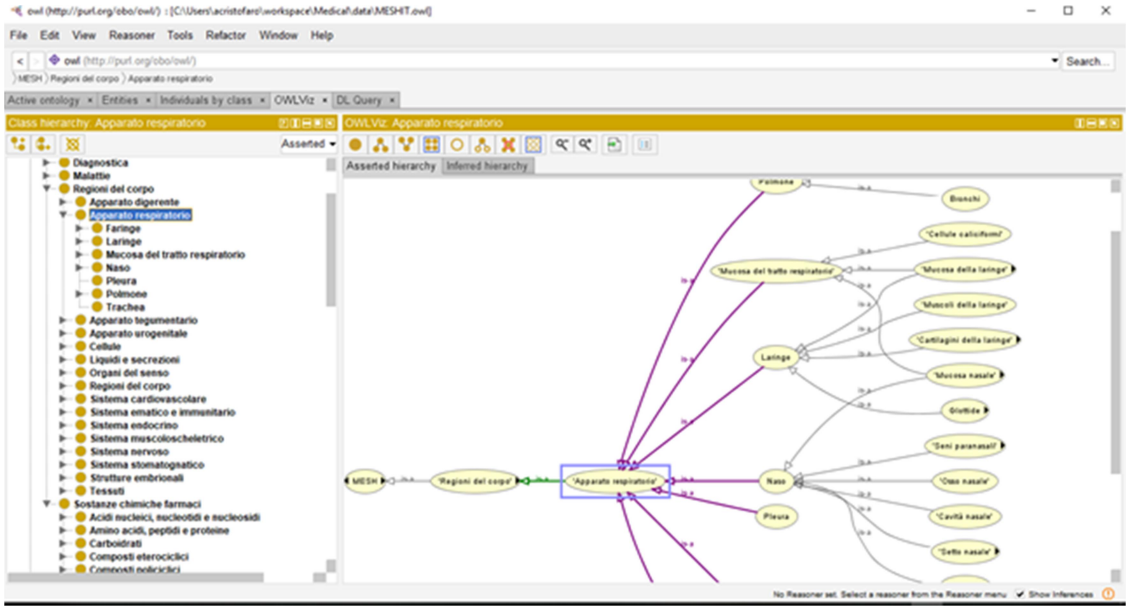
[15] https://protege.stanford.edu/





Come precedentemente espresso questa ontologia è stata utilizzata per la classificazione dei referti in lingua inglese relativi a malati di Covid19.

La seguente figura mostra invece una rappresentazione, sempre mediante Protege, dell'ontologia OWL MeSH che come precedentemente espresso è stata adottata per la classificazione dei referti medici in italiano.



Come è possibile notare osservando la precedente figura, l'ontologia MeSH contiene molte tipologie di concetti che spaziano dalle malattie, fino alle parti del corpo e ai farmaci. Come vedremo in dettaglio quando andremo a descrivere nello specifico i risultati della sperimentazione questa vastità di tipologie di concetti sarà estremamente rilevante nella classificazione dei referti.

Dalle immagini rappresentate è possibile evincere non solo le relazioni tra i concetti (nello specifico relazioni di tipo gerarchico) ma anche le regole di inferenza. Queste complesse relazioni semantiche tra i concetti rendono possibile il reperimento efficace dei risultati basato su concetti semanticamente correlati tra loro oltre alla rappresentazione organizzata dei risultati.

4.2.1.2. *Information Extraction (IE)*

Prima di entrare nel merito dello specifico flusso adottato e delle componenti software utilizzate, forniamo nel seguito una breve panoramica delle tecniche di information extraction. I testi espressi in linguaggio naturale, come è stato già espresso nel capitolo introduttivo, mostrano spesso una grande varietà nel modo in cui uno stesso concetto può essere espresso e quindi una evidente ambiguità semantica. In tali contesti, l'estrazione di informazioni all'interno dei testi richiede sistemi che utilizzano in modo preponderante informazioni linguistiche. Esistono due aree principali di sistemi che

affrontano tale problematica: Information Retrieval (IR) e Information Extraction (IE). La prima, IR, può essere utilizzata per determinare quali documenti, presi da una collezione più o meno vasta, sono rilevanti per una query specificata da un utente, questa può essere semplificata come una lista di parole chiave. Ne consegue che le tecniche di IR permettono di determinare i documenti rilevanti in risposta ad una ricerca. Il ruolo delle tecniche di elaborazione del linguaggio naturale (Natural Language Processing, NLP) risultano essere marginali, anche se questo nell'ambito di ricerca è un argomento controverso[16].

Le tecniche di utilizzo nell'IE richiedono invece una comprensione maggiore di ciò che è scritto nei testi. A differenza dell'IR, tale ambito richiede di determinare contenuti rilevanti all'interno dei documenti, quindi non nella loro interezza. Il contenuto rilevante, esportato in un formato specifico, può essere integrato in sistemi intelligenti anche in coordinazione con sistemi IR per migliorare la ricerca ottenendo risultati più pertinenti. Sistemi di risposta automatica a domande (Question Answering, QA) e sistemi automatici riassuntivi (Automatic Summarization, AS) applicano la combinazione delle due tecniche. La prima è una tecnica che permette di ottenere informazioni a valle di una richiesta specifica di un utente esplicitata in linguaggio naturale. Il più famoso prodotto di QA è probabilmente IBM Watson, il quale è riuscito a battere due campioni del gioco americano Jeopardy! nel 2011[17]. La AS dato un testo più o meno ampio ne fornisce una versione riassuntiva che contiene gli elementi principali del testo in esame ma linguisticamente coerente. Per raggiungere l'obiettivo di comprendere in modo profondo il contenuto di un documento testuale, le tecniche NLP da applicare non si

[16] https://pdfs.semanticscholar.org/8721/f2a087ff35318a056a5814ba287a37df0ec8.pdf oppure Strzalkowski, T., Perez-Carballo, J., Karlgren, J., Hulth, A., Tapanainen, P. and Lahtinen, T.(1999a), Natural Language Information Retrieval: TREC-8 Report, In Proceedings of TREC-8, Gaithersburg, Maryland, USA

[17] https://www.youtube.com/watch?v=WFR3lOm_xhE

sono limitate solo al dividere il testo determinando i termini, eventualmente anche pesati in ordine di importanza, ma ad estrarre informazioni. I metodi statistici non sono sufficienti se non combinati con basi di conoscenza specifici, che devono essere definiti a priori. Questo fa sì che le tecniche di IE dipendono dal dominio in cui vengono applicate, non potendo portare un sistema sviluppato per un ambito specifico in un dominio differente (es. giornalistico vs medico vs economico). Per ovviare a tale limitazione si sono sviluppate tecnologie adattive di IE che facendo uso di metodi empirici di NLP aiutano la costruzione di sistemi IE. Lo sviluppo di metodi e tecniche da applicare alla IE è stato nel passato strettamente collegato a conferenze specifiche del MUC (Message Understanding Conference)[18], che sono state indette negli anni '90, inizialmente sponsorizzate dalla marina americana e successivamente dal DARPA (Defense Advanced Research Projects Agency). Nelle conferenze MUC, gruppi differenti di ricercatori si sono affrontati per ottenere il sistema migliore che fosse in grado di estrarre informazioni da testi ristretti in un dominio specifico. Ad ogni conferenza è stato scelto un dominio differente rispetto al precedente per spingere i ricercatori ad affrontare problemi sempre diversi e quindi ad affinare le tecniche.

Giusto per fare alcuni esempi:

- Nel MUC-3 (1991) è stato scelto un dominio di eventi terroristici nell'america latina. Il materiale di addestramento contemplava 1300 documenti mentre altri 300 costituivano l'insieme per il test e la verifica dei risultati.
- Nel MUC-5 (1993) sono stati scelti documenti che trattavano le Joint Ventures, quindi notizie finanziare, e Microelettronica.
- Nel MUC-7 (1998) è stato scelto un dominio giornalistico di notizie riguardanti incidenti aerei nelle seguenti lingue: cinese, giapponese ed inglese.

I risultati che sono stati riportati dagli organizzatori delle conferenze riportano risultati molto elevati, circa il 91% di F-Measure, ma il problema del riconoscimento e

[18] http://www.itl.nist.gov/iaui/894.02/related projects/muc/

l'estrazione delle entità è ancora lontano dall'essere risolto. Per tale motivo sono state indette altre conferenze in cui la problematica è stata resa più complicata includendo non solo testi, ma anche il parlato ed immagini. Anche le etichette da associare alle informazioni sono state rese meno generali.

Oltre alle iniziative americane, anche la comunità europea ha sponsorizzato progetti nel programma denominato LRE (Linguistic Research and Engineering) per lo sviluppo di strumenti da impiegare nell'IE.

Anche l'Italia ha la sua conferenza specifica, EvalIta[19], che tra le altre cose ha gare sull'italiano per la IE (2007[20], 2009[21], 2011[22], 2016[23]).

4.2.1.2.1. Architettura di riferimento

La prima definizione di una architettura base per sistemi di IE è attribuibile a Hobbs[24] ed è definita come una cascata di "trasduttori" o moduli, che ad ogni passo, aggiungono struttura ai documenti ed in alcuni casi filtrano le informazioni estratte in base a regole. Fare riferimento alla seguente figura per una rappresentazione dell'architettura proposta.

---

[19] http://www.evalita.it

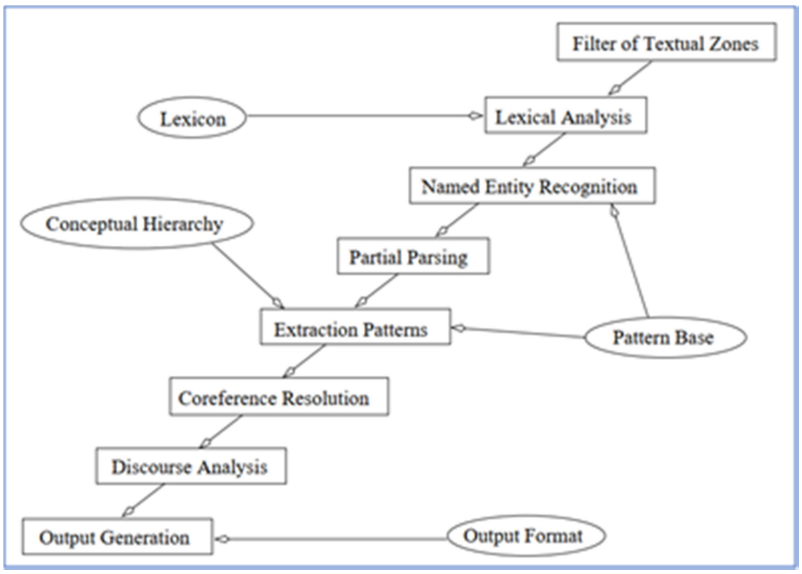
[20] http://www.evalita.it/2007/tasks/ner

[21] http://www.evalita.it/2009/tasks/entity

[22] http://www.evalita.it/2011/tasks/NER

[23] http://www.evalita.it/2016/tasks/neel-it

[24] D. Appelt, J. Bear, J. Hobbs, D. Israel, M. Kameyama, and M. Tyson. Description of the JV-FASTUS System Used for MUC-5. In Proceedings of the 5th Message Understanding Conference (MUC{5), 1993a.



Alcune degli attuali sistemi di IE seguono tale approccio, sostituendo a seconda dell'autore un modulo con un altro. I componenti possono riassunti in elementi macro come segue:

- Modulo di pre-processamento di documenti.
- Parser sintattico dei documenti, parsing in modo completo o parziale.
- Interpretatore semantico, per generare una forma logica (una espressione con una sintassi specifica che definisce una semantica precisa. Per esempio, se A=>B ed A è vero allora B è vero: "tutti gli uomini sono mortali", "Socrate è un uomo" allora "Socrate è mortale") o un template che viene riempito con gli elementi riconosciuti nel testo.
- Analizzatore del discorso, per legare i vari elementi semantici presenti nelle frasi di un testo (anaphora resolution è una di esse).
- Generatore finale, per trasformare le informazioni derivate nel formato di uscita finale.

4.2.1.2.2. Modulo di pre-processamento di documenti

Tale componente può essere composto di diversi moduli:

- "**text zoner**", riconoscono specifiche zone del testo quali: titolo, finale, bibliografia, ecc.

- "**segmenter**" (o "splitter"), dividono il testo in segmenti che possono coincidere con i paragrafi di un testo o con frasi ("sentence splitter").
- "**tokenizer**", dividono il testo in elementi lessicali (parole).
- "**language guesser**", riconoscono la lingua in cui un documento è scritto tramite compionamento di parole.
- "**lexical analyzer**", procedono con l'analisi lessicale del testo e può includere sia l'analisi morfosintattica che il riconoscimento di entità nominate (Named Entity) con relativa disambiguazione.
- "**POS tagger**", etichetta ogni parola del testo con una etichetta relativa all'analisi grammaticale ("nomi", "congiunzioni", "articoli", ecc).
- "**stemmer**", è un normalizzatore che fornisce per ogni parola coniugata la sua forma base senza suffisso (o prefisso). Esempio: "studies" diventa "studi" avendo tolto il suffisso.
- "**lemmatizer**", è un normalizzatore che fornisce per ogni parola conuigata il suo lemma di riferimento. Esempio: "studies" diventa "study" che è il verbo corrispondente all'infinito.

I sistemi di IE per il pre-processamento dei dei documenti, fanno uso di risorse linguistiche o di componenti indipendenti dal dominio. Molto importante invece è il modulo di riconoscimento delle entità nominate (Named Entity Recognizer, NER), questi possono far uso di automi a stati finiti (Finite State Transducer) o ricerca in dizionari (dictionary lookup).

4.2.1.2.3. Named Entity Recognition

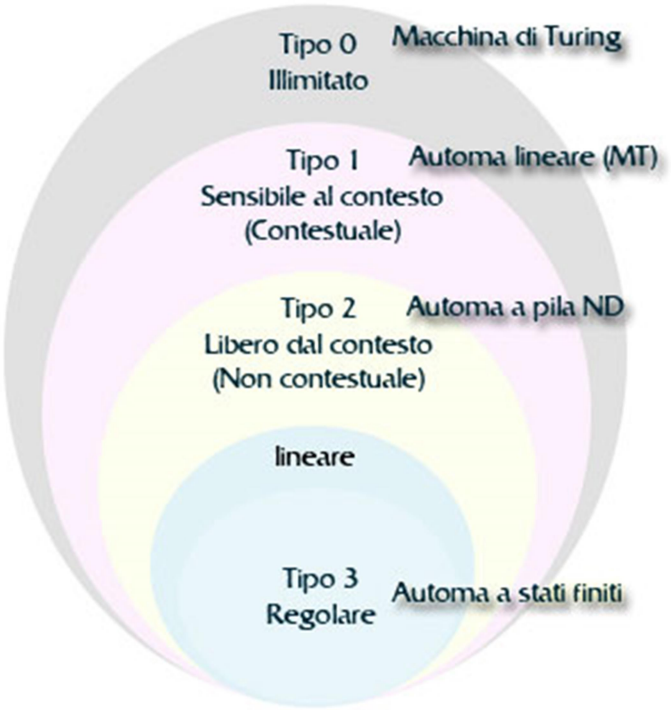
Come introdotto nella sezione precedente, un componente per il riconoscimento delle entità nominate (NER) è cruciale per un sistema di IE. Lo scopo di tali componenti è quello di identificare all'interno dei testi, le entità nominate cioè parole che assumono uno specifico significato nel testo quali: persone, luoghi, organizzazioni, ecc. Se come esempio si considera la seguente frase: "Sergio Mattarella è in Italia il Presidente della Repubblica.", le entità sono:

- "Sergio Mattarella", persona.
- "Italia", luogo.

- "Presidente della Repubblica", persona.

È di comune accordo nella comunità scientifica di considerare entità nominate anche le quantità siano essere monetarie o temporali.

Esistono diverse metodologie per il NER, ci sono sia approcci basate a regole grammaticali, che modelli statistici basati ad esempio al campo dell'apprendimento automatico. I sistemi basati su regole, possono ottenere risultati molto precisi a scapito della copertura del numero di fenomeni riconoscibili. Inoltre, diventano non manutenibili all'aumentare del numero di regole da utilizzare, ed estremamente costose dovendo essere scritte manualmente da esperti di settore. C'è inoltre un vincolo intrinseco nel metodo e riguarda le logiche formali che soggiacciono il modello scelto. Se si considera infatti la famosa gerarchia di Chomsky, mostrata nella figura seguente, se l'automa scelto è di tipo a stati finiti (equivalente per intenderci ad una regular expression) è relativamente semplice da costruire ma la possibilità di riconoscere correttamente le entità all'interno di un testo di per sé dipendente dal contesto è limitata. In un ambito ambiguo come la lingua, potrebbe infatti capitare di dover riconoscere sia persone e luoghi, e se non si considera il contesto, è impossibile discernere quando viene usata la parola "Washington" come persona (il primo Presidente degli Stati Uniti) dalla capitale degli Stati Uniti.



Si tenga presente che all'altro estremo della gerarchia di Chomsky c'è la Macchina di Turing, cioè il modello teorico che soggiace i computer odierni, e la grammatica naturale si trova in un livello intermedio tra il "tipo 1" e il "tipo 2". Quando fu definita tale gerarchia, si pensò che fosse possibile definire una grammatica formale che potesse descrivere la grammatica naturale vera a propria, diversi studiosi hanno provato ad ottenere ciò ma non esiste ancora nessuno che ci sia riuscito e si hanno dubbi che questa possa essere raggiunta in un modo o nell'altro. E comunque anche se si ottenesse sarebbe estremamente complessa ed in alcuni casi non utilizzabile a causa dell'ambiguità della lingua. Il classico esempio è: "la vecchia porta la sbarra". La parola "porta" nella suddetta frase è "verbo" o "nome"? Non esiste una risposta giusta o sbagliata, sono purtroppo vere entrambi e nemmeno un essere umano riesce a capirlo se non utilizzando altri sensi, per esempio il tono di voce o il contesto visivo in cui si trova, e comunque non è sempre possibile.

I sistemi che invece fanno uso dell'apprendimento automatico, invece affrontano il problema in modo diverso, si focalizzano sull'analizzare un dataset, scritto da un essere umano, apprendendo dei modelli che poi potranno essere riutilizzati in altri casi. Tali

approcci però hanno bisogno di un grande quantitativo di materiale di addestramento e purtroppo non possono essere utilizzati in contesti differenti da quelli per cui sono stati pensati.

Negli ultimi hanno stanno prendendo piede piattaforme che applicano una logica di crowd contribution. In tali piattaforme una moltitudine di contributori, possono accedere e partecipare all'annotazione collettiva di documenti, utilizzando sistemi di etichette riferenti a categorie di Wikipedia.

5. Conclusione

In svariati modi abbiamo mostrato come la piattaforma Digital Future diventa uno strumento essenziale nelle attività di analisi dei dati nell'ambito Healthcare.

I metodi di analisi predittiva della piattaforma Digital Future apprendono dai dati storici e fanno predizioni accurate sui risultati. Elaborano i dati del paziente, danno un senso alle note cliniche, trovano le correlazioni, le associazioni di sintomi, gli antecedenti familiari, le abitudini, le malattie e poi fanno previsioni. Gli impatti di alcuni fattori biomedici come la struttura del genoma o le variabili cliniche sono presi in considerazione per prevedere l'evoluzione di alcune malattie. I casi più comuni includono la prognosi del progresso della malattia o la prevenzione per ridurre il rischio e gli esiti negativi. Il principale vantaggio è il miglioramento della qualità della vita dei pazienti e la qualità delle condizioni di lavoro per i medici.

Inoltre, i dispositivi indossabili stanno diventando rapidamente onnipresenti. Oltre a creare un accessorio interessante, incoraggiano la gestione della salute personale nelle persone. Registrano importanti letture della salute come pressione sanguigna, battito cardiaco, ritmo del sonno, polso ecc.

Per gli anziani, in particolare, i dispositivi indossabili sono di grande aiuto. Aiutano i membri della famiglia a rimanere aggiornati sulla salute dei pazienti e ad informare immediatamente lo staff medico in caso di emergenza. I dispositivi indossabili possono essere collegati alla piattaforma Digital Future e continuano a generare e archiviare volumi di dati dei pazienti a cui è possibile accedere quando necessario.

Utilizzando l'intelligenza artificiale, l'apprendimento automatico e i big data, i tenant possono analizzare queste informazioni non elaborate dai dispositivi indossabili per ricavarne informazioni significative. Attraverso modelli analitici avanzati, possono osservare variazioni nella salute del paziente e rilevare un disturbo o un sintomo relativo. Questo aiuta i medici a prevedere un possibile problema di salute e fornire cure preventive.

Messe insieme, queste disposizioni basate sui dati possono far posto a costi operativi ridotti che si traducono in minori costi sanitari per i pazienti con maggiore soddisfazione. Ciò sarà determinante per ottimizzare l'assistenza e l'esperienza del paziente.

6. Bibliografia e Sitografia

1. MONITOR MULTIPARAMETRI PC-300. In: Gimaitaly [Internet]. Available: https://www.gimaitaly.com/prodotti.asp?sku=35162&dept_selected=610&dept_id=610

2. NuvoAir. NuvoAir - Measure your lung health at home with Air Next Spirometer [Internet]. Available: https://www.nuvoair.com/

3. Mi Italy [Internet]. Available: https://www.mi.com/it/mi-band-3/